To appear in the International Journal of Pattern Recognition and Artificial Intelligence, special issue on Speech Recognition for different languages, Spring 1994.

Speech-to-Text Conversion in French

J.L. Gauvain, L.F. Lamel, G. Adda, and J. Mariani

LIMSI-CNRS
BP 133
91403 Orsay cedex, FRANCE

Keywords: Speech-to-text conversion, continuous speech recognition, speaker-independent recognition, vocabulary-independent recognition, phone recognition, hidden Markov models, language identification,

Abstract

Speech-to-text conversion of French necessitates that both the acoustic level recognition and language modeling be tailored to the French language. Work in this area was initiated at LIMSI over 10 years ago. In this paper a summary of the ongoing research in this direction is presented. Included are studies on distributional properties of French text materials; problems specific to speech-to-text conversion particular to French; studies in phoneme-to-grapheme conversion, for continuous, error-free phonemic strings; past work on isolated-word speech-to-text conversion; and more recent work on continuous-speech speech-to-text conversion. Also demonstrated is the use of phone recognition for both language and speaker identification.

The continuous speech-to-text conversion for French is based on a speaker-independent, vocabulary-independent recognizer. In this paper phone recognition and word recognition results are reported evaluating this recognizer on read speech taken from the BREF corpus. The recognizer was trained on over 4 hours of speech from 57 speakers, and tested on sentences from an independent set of 19 speakers. A phone accuracy of 78.7% was obtained using a set of 35 phones. The word accuracy was 88% for a 1139 word lexicon and 86% for a 2716 word lexicon, with a word pair grammar with respective perplexities of 100 and 160. Using a bigram grammar word accuracies of 85.5% and 81.7% were obtained with 5K and 20K word vocabularies, with respective perplexities of 122 and 205.

1 Introduction

At LIMSI, the idea of realizing a voice-activated typewriter with a very large dictionary was initiated in the late 1970s. The first experiment on this topic concerned the phoneme-to-grapheme conversion of error-free continuous phoneme strings. Phoneme-to-grapheme conversion consists of segmenting the phoneme string into words and then generating the correct orthographic translation of those words. The initial step of segmentation used a

simple heuristic of choosing the solution with the smallest number of words for a given sentence[34]. Evaluated on a small text corpus containing 1796 words, using a lexicon of 20,000 baseforms, the error rate in segmentation was suprisingly low, only 20%, with most of the errors due to liaisons. This first step in the direction of a voice-activated typewriter has since been followed by more extensive efforts.

This work was extended to convert phoneme strings containing simulated errors[4] and the methodology was adapted to stenotype-to-grapheme conversion[1] using statistical language models trained on text corpora. In the framework of the ESPRIT project 860 "Linguistic Analysis of the European Languages," LIMSI's approach to language modeling was compared with other closely related approaches, on 7 different European languages[6]. The 4-year project was followed by the ESPRIT Polyglot project, which had the goal of designing speech-to-text and text-to-speech systems for each of the same 7 languages.

Work in speech recognition at LIMSI also began in the 1970s, continuing to the first French single-board isolated-word speech recognizer, *Moise*, and the first single-board connected-word recognizer, *Mozart* [19, 14] which was able to recognize a vocabulary of about 100 words. The link between acoustic recognition and language modeling was made simultaneously with the development of the *Hamlet*, 2000-word, speaker-dependent (SD) isolated-word (IW) dictation system[36], and with the 7000-word, SD IW dictation system developed within the *Amadeus* speech recognition project[13]. The recognizer in the *Amadeus* project was developed around a specialized DTW chip (μPCD)[44, 45] that has been designed at LIMSI, in collaboration with the Bull and the Vecsys companies. Acoustic recognition was first demonstrated with a chip emulator in March 1987, and a complete dictation system using the chip itself was demonstrated in spring 1988.

Presently, the primary research efforts in speech recognition are directed at the dictation task and a dialog project. For both applications, a speaker-independent (SI), vocabulary-independent (VI), phone recognizer is being developed, so as to be easily adaptable to various tasks.

In the dictation task, the BREF corpus [18, 30], described in more detail in Section 5.2, is used. The immediate goal is to work with read speech material from a large number of speakers, so as to be able to build base acoustic models which can be augmented and adapted to specific speakers or tasks. This work also allows many aspects of language modeling to be addressed under more "semi-controlled conditions," than those found in spontaneous dictation. Additionally, it is much easier to collect read-text material than spontaneous dictations.

An ongoing dialog project is oriented toward Air-Traffic Controller training, in collaboration with the Centre d'Etudes de la Navigation Aérienne. Currently, the student training sessions are limited by the availability of the human instructor who plays the role of a pilot. The goal is to replace the instructor by a spoken dialog system. This allows for more availability of the system, and may force the student to adhere to the pre-defined phraseology, the learning of which is part of training. The dialog system is built around the *Amadeus* speech recognizer and an associated synthesis module.

The remainder of this paper is as follows. First, some of the distributional properties of French, most of which are gathered from the study of large text corpora, are presented. In Section 3 some of the problems encountered in speech-to-text conversion are addressed, highlighting those problems specific to French. Section 4 describes an approach to isolated-word speech-to-text conversion, and Section 5 presents more recent efforts using continuous speech. This section includes phonetic and word recognition, as well as some issues in language modeling.

2 Distributional properties of French

In order to be effective in speech-to-text conversion, it is necessary to determine and account for the distributional properties of the language. For practical reasons, it will probably remain impossible to obtain text transcriptions of sufficient spoken utterances for analysis and language modeling. Therefore, large written text corpora serve as a basis for analysis and modeling the distributional properties of spoken texts. As long as applications such as text dictation remain of interest, the models thus obtained should be relatively reflective of the task.

In this section the analysis of a large corpus of text material [18, 30] taken from the French newspaper *Le Monde* is presented, along with some comparative data taken from a smaller text corpus of Senate transcripts. The source text materials consisted of three months of *Le Monde*, representing about 5 million words of text and 1.2 million words of Senate transcriptions. After cleaning up the newspaper text so as to eliminate incomplete sentences and to correct formatting errors, 4.2 million words remained. The "lost" text was roughly 50% header information and 50% textual errors.

The distributional properties of the texts were determined by counting the occurrences of sentence, word, and subword units. At the sentence level, counts were made of sentence types and lengths. At the word level, the number of distinct words and their word frequencies were counted. Subword units counted include syllables, dissyllables, phones, diphones, and triphones.

2.1 Text Analysis

Each sentence was phoneticized using grapheme-to-phoneme rules[43], and erroneous pronunciations were hand-located[1] and corrected using an exceptions dictionary. The most common mispronunciations were foreign words and names, and acronyms. Also, each punctuation mark was replaced by a silence "phone." The set of phone labels used in grapheme-to-phoneme rules is given in Table 1. Although certain speakers of French make the distinction between the vowel in the words "p<u>a</u>tte" and "p<u>â</u>te", and the nasal vowels in the words "br<u>in</u>" and "br<u>un</u>", they have been collapsed together as the majority of speakers do not reliably distinguish between them.

2.1.1 Sentence types:

Sentences were classified as declarative, interrogative and exclamative types, or as more complex formulations which included ellipses, parenthetic expressions, and/or quotations. Table 2 shows the distribution of sentences in *Le Monde* according to type, and shows for each type the minimum, average, and maximum sentence lengths. Simple sentences contain no internal punctuation markers other than comma, and no embedded parenthetic expressions or quotations. Conversely, complex sentences contain at least one of these. For the sentence lengths, the counts are for split quotations. The final part of the table gives the percentage of sentences containing numbers, acronyms, quotations (entire and split into individual sentences), or parenthetic expressions.

A conceptual problem was found while counting sentence types, a priori, a simple task: what should be done with end-of-sentence punctuation marks found within parenthetics or

[1] Since this is such a labor-intensive procedure, corrections were made only for words occurring more than 20 times in the text.

Phone	*Example*	*Phone*	*Example*
Vowels		Consonants	
i	l<u>i</u>t	s	<u>s</u>ot
e	bl<u>é</u>	z	<u>z</u>èbre
E	s<u>e</u>l	S	<u>ch</u>at
y	s<u>u</u>c	Z	<u>j</u>our
X	l<u>eu</u>r	f	<u>f</u>ou
x	p<u>e</u>tit	v	<u>v</u>in
@	f<u>eu</u>	m	<u>m</u>otte
a	p<u>a</u>tte, p<u>â</u>te	n	<u>n</u>ote
c	s<u>o</u>l	N	di<u>gn</u>e
o	s<u>au</u>le	l	<u>l</u>a
u	f<u>ou</u>	r	<u>r</u>ond
Nasal Vowels		p	<u>p</u>ont
I	br<u>in</u>, br<u>un</u>	b	<u>b</u>on
A	ch<u>an</u>t	t	<u>t</u>on
O	b<u>on</u>	d	<u>d</u>on
Semivowels		k	<u>c</u>ou
h	l<u>ui</u>	g	<u>g</u>ond
w	<u>oui</u>	·	silence
j	<u>y</u>ole		

Table 1: The 35 phone symbol set.

quotations? The analysis was performed two ways, ignoring and counting these marks. However, in sentence selection, it was decided to ignore parenthetic expressions as they are often too disjoint from the text, and to divide sentences within a long quotation into single, quoted sentences. This decision was made because sentences containing complex quotations could be quite long - over 500 sentences were found having more than 100 words each! While 12% of the quotations were only a single word and another 25% were 2-5 words long, the average length for a single quotation was 11 words. In contrast, parenthetics were typically short: over 75% had fewer than 5 words and the average length was 4 words.

2.1.2 Word and subword units:

Word and subword units were counted in the phonemicized, syllabified text. Punctuation markers were considered to be non-verbalized, and therefore were not counted as words. Table 3 summarizes the counts for the different units for the complete text of *Le Monde* and the Senat. Counts made on only 10% of the text of *Le Monde* showed almost identical distributional properties.

In the 167,359 sentences, there were almost 4.2 million words, with over 90,000 orthographically distinct. To find the number of phonemic words, the grapheme-to-phoneme mapping was redone without the liaison rules, so as to avoid the ambiguity in word segmentation introduced by liaison. There were 64,000 phonemically distinct words, almost 30% less than the number of orthographically distinct words, giving a measure of the number of homophones in French. In order to know if the percent of homophones was dependent upon the vocabulary size, the percent homophones in 2000 and 10,000 most common words were determined, and also found to be roughly 30%. The dissyllable is defined from the midpoint of one vowel to

Sentence Type	*Percent*	*Number of Words*		
		Ave	*Min*	*Max*
Declarative	95	23	1	222
Interrogative	3.8	15	1	191
Exclamatory	1.2	13	1	104
Simple Sentences	57	19	1	191
Complex Sentences	43	33	3	222
Numbers	22	30	1	165
Acronyms	11	-	-	-
Split Quotations	27	26	2	213
Quotations	22	34	2	>400
Parenthetic	11	35	2	>100

Table 2: Sentence types and lengths.

the midpoint of the next vowel, and therefore contains all the intervening consonants. This unit has been successfully used for speech recognition and speech synthesis in French[50], in part because French vowels are acoustically relatively stable over time.

Unit	*Le Monde*	*Senat*
#sentences	167,359	64,613
#words (total)	4,244,810	1,137,928
#orthographically distinct	92,185	26,807
#phonemically distinct	63,981	
#syllables (total)	6,903,017	1,956,423
#distinct syllables	9,571	
#distinct dissyllables	37,636	
#phones (total)	16,416,738	4,737,578
#distinct phones	35	35
#distinct diphones	1,160	1,105
#distinct triphones	25,999	17,079

Table 3: Distributional properties of word and subword units.

On the average, there were 2.3 phones/syllable, 3.2 phones/dissyllable (including both vowels), and 3.7 phones/word. The most common phone was /r/, accounting for 8.0% and 7.9% of all phone occurrences in *Le Monde* and Senat, respectively. Most of the possible diphones were found to exist (1160 out of 1225, taking into account the silence "phone"), as were 60% of the possible triphones. Some of these gaps are truly indicative of the French language, while others may be due to insufficient data or the grapheme-to-phoneme rules. However, the number of triphones may actually be elevated, relative to "traditional French", since there are so many foreign words (mostly names) in the text source.

Figure 1 shows plots of the frequency of occurrence for the word and subword units in percentages. Part (a) has curves for words, syllables, and phones, and part (b) has curves for dissyllables, triphones, diphones, and phones. The units have been separated as such since

Figure 1: Frequency of occurrence for word and subword units.

words, syllables, and phones have no constraints internal to the unit itself restricting which units may follow, whereas the units in part (b) have internal constraints limiting the possible following units. Phones are shown in both for comparison as the basic unit.

Less than 20% of the distinct words account for over 95% of all word occurrences. In fact, 40% (about 35,000 words) occurred only once in the text, and 60% of the words appeared at most 3 times. This effect is even more pronounced for syllables, where the roughly 20% most common syllables account for 98% of all syllable occurrences. Almost 80% of the text is covered by only the most frequent 232 (20%) diphones. 20% of the triphones and dissyllables cover over 90% and 95% of the text, respectively.

Figure 2: Percentage of sentences covered as a function of unit.

But perhaps more interesting is the opposite question: given that 40% of the words only occurred once in the text, how many sentences can be pronounced if these words are eliminated? The curves shown in Figure 2 illustrate the percentage of sentences covered as a function of the percentage of word or subword unit. The curve for phones is very gradual - with 80% of the phones, only 10% of the sentences can be covered. For words, however, over 80% of the sentences are covered using only 60% of the distinct words, effectively eliminating all of the single occurrence words. The effect is even stronger for syllables: roughly 40% of the syllables cover over 90% of the sentences. Curves are shown for phones, diphones, triphones, and dissyllables in Figure 2b.

2.2 Entropy

In order to assess the relative importance of the word and subword units, the entropy of corresponding Markov sources were calculated. The probabilities used for each source are shown in Table 4a, where w_i, s_i, v_i, and a_i are respectively a word, syllable, vowel, and phone, and c_k is a string of consonants. A memoryless source was used to model the phone, word, and syllable sources. The diphone and dissyllable models were first order Markov sources, and the triphone model was a second order Markov source. All probabilities were estimated using frequency counts on the entire text.

(a) *Unit*	*order 0*	*order 1*	*order 2*
phonemic words	$p(w_i)$		
syllables	$p(s_i)$		
dissyllables	$p(v_i)$	$p(c_k, v_j \mid v_i)$	
phones	$p(a_i)$		
diphones	$p(a_i)$	$p(a_j \mid a_i)$	
triphones	$p(a_i)$	$p(a_j \mid a_i)$	$p(a_k \mid a_i, a_j)$

(b) *Unit*	*#Distinct units*	*Entropy (b/ph)*	*Model I (b/ph)*
phonemic words	63,981	2.67	2.46
syllables	9,571	3.61	1.51
dissyllables	37,636	3.55	1.57
phones	35	4.72	0.40
diphones	1,160	3.92	1.21
triphones	25,999	3.40	1.72

Table 4: Markov sources: (a) model probabilities and (b) estimated entropies.

Table 4b summarizes the results of the models in bits/phone. The lowest entropies are found for the word and triphone sources, indicating that their models store the most information. Compared to the memoryless, equally probable 35 phone source, the information stored in the models is 2.46 and 1.72 b/ph, respectively.

3 Problems in speech-to-text conversion

Phoneme-to-grapheme conversion of French seems to be more difficult than in other languages, due to the large number of homophones. Starting with a source dictionary of 22,000 baseforms results in a full-form lexicon of about 162,900 graphemic words. Grapheme-to-phoneme translation of those words produces about 90,000 distinct phonemic forms, indicating that for a large full-form lexicon, a phonemic word corresponds to, on the average, 1.8 different graphemic words. Table 5 shows some approximate full form counts for a baseform lexicon of 22,000 words. For comparison, there are roughly 3% homophones in the DARPA Resource Managment lexicon[42] , less than 2% for the DARPA TIMIT lexicon[31, 11], and under 5% in the MIT Pocket lexicon[52].

In fact, the main problem arises from verb conjugation. A single verb has on average 40 forms. Among these, there are as many as three different spellings for each pronunciation.

Another source of homophones is that the mark of plurals (an -s at the end of the word) for most substantives, most adjectives, and all the past participles, is never pronounced in isolation, and only sometimes pronounced in fluent speech. Similarly, the mark of the feminine form (-e at the end of the word) for some substantives, most of the adjectives and the past participles, is never pronounced.

In addition, there are the more typical "word" homophones, such as the demonstrative adjective *ces (those)* and the possessive adjective *ses (his)*, which have the same pronunciation /se/. Some examples of the different types of homophones are given in Figure 3.

	% Words	*# Words*	*# Forms/Word*	*# Forms*
Verbs	14%	3,100	40	124,000
Substantives	56%	12,300	2	24,600
Adjectives	23%	5,100	2.5	12,800
Adverbs and others	7%	1,500	1	1,500
Total	100%	22,000	(avg.) 7.3	162,900

Table 5: Full-forms derived from a dictionary with 22,000 baseforms.

Verbs:
/kas/ casse, casses, cassent (break)
Substantives (Masculine/Feminine):
/ami/ ami (friend (he)), amie (friend (she))
Substantives (Singular/Plural):
/tas/ tasse, tasses (cup, cups)
Adjectives (Masculine/feminine):
/ene/ aîné (older masc.), aînée (older fem.)
Adjectives (Singular/plural):
/grAd/ grande, grandes (big)
Past Participles:
/kase/ cassé, cassés, cassée, cassées (broken)

Figure 3: Examples of common homophones.

Considering now the case of continuous speech, the problem of segmenting the continuous phoneme string into words seems to be especially difficult in French. In experiments on a simple sentence containing 9 phonemes, "J'ai mal au pied." (My foot hurts.), with the 162,900 word full-form lexicon, more than 32,000 possible transcriptions (segmentations and orthographic translations) were obtained at the lexical level. As shown in Figure 4(a) even using phonological rules, syntax, and semantics still leaves two acceptable sentences that require a pragmatic analysis in order to get the right graphemic transcription. Another example, shown in Figure 4(b), gives the possible analyses of the phrase "un murmure de mécontentement". This example illustrates both the complexity of the problem and the power of the syntactic constraints. Lexical access using a full-form lexicon with over 300,000 entries yields 340 possible word segmentations. This expands to over 2 million possible phrases when all the combinations are considered. Syntactic constraints including form agreement reduce

the set to 6 possibilities, all of which are semantically plausible.

Text string: J'ai mal au pied.
Phonemic string: /Zemalopje/
Possible segmentations at the lexical level:
32,000 possibilities
Geai mâle au pied
Geais ma lot pieds
J'hait mât l'eau piller
J'aime allo pillé
J'aimes allo pillé
Jet malles hop y est
Gemme halles hopi et
....
Possible segmentations with phonology, syntax, and semantics:
J'ai mal au pied. (My foot hurts.)
J'ai mal aux pieds (My feet hurt.)

(a)

Text string: un murmure de mécontentement
Phonemic string: /ImyrmyrdxmekOtAtmA/
Lexical access: 340 possible word segmentations
2,419,620 phrases
Syntactic analysis: 6 possible phrases
- un murmure de mécontentement
- un murmure de mécontentes ment
- un murmure de mes contentements
- un mur mûr de mécontentement
- un mur mûr de mécontentes ment
- un mur mûr de mes contentements

(b)

Figure 4: Lexical hypotheses from an error-free phonemic transcription.

In French one must also deal with "liaison", the links made between words. These are phonemes that are pronounced at the junctions between two words, but would not be pronounced at the end of the first word, or at the beginning of the second one, if the words were pronounced in isolation. For example, the word sequence "les amis" (the friends) is pronounced /lezami/, where the word pronunciations in isolation would be /le/ and /ami/. Another more complicated form of liaison is the insertion of /t/, in certain inverted verb forms. Instead of forming the question "A il ...", the written and spoken form is "A-t-il ...". In certain cases this liaison is the only indication to distinguish between the singular and plural forms of a word. This is true of the phrases "Il aime le pain." (He likes bread.) and "Ils aiment le pain." (They like bread).

Another problem is the optional pronunciation of mute-e. For example, the word *devenu* can be produced with 2 or 3 syllables: /dxvnu/ or /dxvxnu/. The same phenomena can also

occur across words: *beaucoup de gens* may be pronounced as /bokudxZA/ or as /bokudZA/. This problem of schwa-deletion is also found in English, however the phonemic environments are somewhat different. Additionally, there are situations where the word-final mute-e is pronounced. This effect is to some extent context and dialect dependent. Speakers from the south of France typically enunciate the mute-e, whereas speakers from the Parisian area will usually leave it out.

A final problem which is mentioned only briefly here is with apostrophe, where the final vowel of certain words can be deleted when the next word begins with a vowel. In the written form, this results in words like *l'enfant, c'est, n'a, s'amuser ....* This problem is discussed in more detail in Section 4.4.

4 Isolated-Word speech-to-text

In this section our efforts in isolated-word speech-to-text conversion are described. The goal of this work was to integrate the necessary components of an isolated word, speaker-dependent Voice Activated Typewriter (VAT) on a stand-alone personal computer[36, 13]. The target vocabulary size was several thousand words. The language model was given by bigrams and trigrams of grammatical categories[3, 8], where the probabilities were computed by counting the occurrences in the training text material. Recognition was a two step process. First, a fast match to select a small subset of the lexicon, then a detailed, DTW-based word match was performed that gives the list of word-candidates with their recognition score. On average, the fast match returned about 2% of the lexical entries.

Two systems were developed, both using specialized hardware for signal processing. The first, having a vocabulary of 2,500 words ran on an IBM PC workstation. The second ran also on a IBM PC, but took advantage of the (μPCD) custom VLSI search processor[44] to perform DTW operations. This processor was been designed to be used in applications using pattern matching operations (Speech and Character Recognition, Stereovision, Scene Analysis, Operational Research...). The processor is fully programmable and can support isolated-word and connected-word recognition algorithms using DTW or HMM approaches. Using the DTW approach, it can perform recognition with a full search on a vocabulary of 1,000 words in isolation, or 300 words spoken continuously, in real time. By adding a fast-match algorithm also supported by the processor, it allows real-time recognition of a vocabulary of 7,000 words in isolation. The vocabulary size can easily be extended by multiplying the number of such processors. A single IBM PC board can hold up to 16 processors.

4.1 Language Modeling

In this early work two small applications were explored. The first was related to dictating a research report in the field of speech technology in French. The training data consisted of an existing 20 page research report containing about 15,000 words. There were on the order of 2500 distinct graphemic forms and 2000 phonemic forms. The second application was oriented to dictation of more general French. The vocabulary was defined by a French textbook for foreigners, "Le Mot et L'Idée,"containing about 40,000 words of text. There were 6,700 distinct graphemic forms and 5,100 distinct phonemic forms in the text.

For these applications it was decided to pronounce all punctuation markers as a word, and to speak numbers as digit strings, unless they were included in a word. Considering the

peculiarities of French presented in the previous section, it was decided to pronounce the apostrophe as a word, and to leave unsaid the liaisons between words.

Label	*Grammatical Class*	*Example*
PONC	punctuation	,
NM	noun masculine	animal
NMP	noun masculine plural	animaux
NF	noun feminine	fleur
NFP	noun feminine plural	fleurs
AM	article masculine	le
AF	article feminine	la
AP	article plural	les
INF	verb infinitive	chanter
VPP	verb past participle	perdu
AJM	adjective masculine	beau
AJMP	adjective masculine plural	beaux
AJF	adjective feminine	belle
AJMF	adjective feminine plural	belles
PN	pronoun	je
PO	possessive	me
NUM	number	dix
NPR	proper name	Daniel

Table 6: Some of the grammatical classes used in the general dictation task.

In order to provide constraints, sets of grammatical categories were defined. For the general French dictation task a set of 59 grammatical classes were used. Some of these classes are given in Table 6. The nouns and adjectives have been subdivided into four classes to handle the masculine/feminine and singular/plural distinctions. For the research report dictation an extended set of 160 grammatical classes were defined. These classes, which were closely related to the categories used by other authors[8], were obtained from 55 basic categories, by adding gender or number information, such as the classes "substantive masculine singular" or "article feminine singular". The grammar takes into account the fact that an apostrophe must be followed by a word beginning by a vowel, as well as other rules such as that the possessive adjective "mon" cannot be followed by a feminine word beginning with a consonant.

Language model training for the general dictation task consisted of building n-gram grammars for the grammatical classes on the entire text book. In contrast, an incremental training method was used for report dictation. In this scheme, each successive page of the text report was analyzed using the language model built from the previous pages (for the first page, it started from scratch). Each word in the text was looked up in the lexicon. If it was found, its phonemic representation and grammatical category were specified. If not, its phonemic representation was obtained by grapheme-to-phoneme conversion, and its grammatical category was inferred inductively using a stochastic syntactic parsing method. The result of this analysis was a "verticalized" text[6], where each graphemic word of the text was followed by its phonemic translation, its grammatical category, and the type of inference (lexical or syntactic) used to get the information. This page of text was then manually corrected, and

used to update the lexicon and the syntax, that was then used to process the next page.

4.2 Acoustic Training

The training speech data were recorded in a relatively quiet office environment. During the training phase, all the words of the phonemic vocabulary were pronounced once. In the hardware implementation, the speech signal was filtered at 4.8 kHz, and sampled at 10 kHz. Eight Mel-frequency scale cepstral components were computed every 12.5 ms. A non-linear time compression algorithm[19, 23] was used to compress the steady-state portions of the signal. In order to reduce the size of the reference templates, vector quantization was applied.

4.3 Evaluation

The system has been tested on the vocabulary of the textbook in French for foreigners. It has 5,127 phonemic words, corresponding to 6,700 graphemic words. On a 1000-word text dictated by one speaker, a phonemic word recognition rate of 91% was obtained. This increased to 99% correct phonemic word recognition using the language model. Recognition of the graphemic words was 92.5%, with 75 errors. All tests were made on text data that were used for building the language model. The average recognition time for a word was 480 ms.

An example dictation output for this application is given in Figure 5. The example text was dictated in one continuous session and corresponds to a page in the book. In general the errors made by the system were confusions with another acoustically similar word. The word "hommes" is seen to be consistently misrecognized as "pommes", suggesting that the model for this word was poor. The constraints provided by the language model could not differentiate amongst "il reste" and "ils restent" or "quelque élève" and "quelques élèves", both of which are homophones when spoken in isolation. Apparently the singular form was more common in the training data, and therefore was selected here. These word pairs are even homophones in continuous speech, and cannot be disambiguated without additional semantic information, as given by the "les enfants" in the preceding sentence. The confusion between "fonds" (business) and "fond" (rear) also cannot be eliminated without semantic information.

An example dictation output for the report application is given in Figure 6 for the phrase "se sont portés vers les problème relatifs" (have been conducted on the problems related). The hypothesized list of candidates for each word are shown as a list. The recognized words are shown in bold face.

For this task, most of the recognition errors were made on 1 or 2-syllable words. As the shortest words, which seem to be the most difficult to recognize, are also the most frequent ones, it is expected that word recognition rates on text dictation are worse than error rates reported on word lists. However, since these short words are very common, they are also well represented in the language model. Therefore the language model can greatly help in correcting the "acoustic" recognition errors made on these short words. A related effect was that the recognition rate did not vary when the size of the lexicon was increased from 1500 words to 2000 words. This may be due to the incremental approach used to build the lexicon: since the shortest, most error-prone words are rapidly included in the lexicon, extending the lexicon tends to add longer, less confusable words.

L' activité corporelle.

1. Les pommes *(hommes)* et les animaux peuvent remuer, se mouvoir, se donner du mouvement. Les pommes *(hommes)* sont capables de faire des gestes de la tête et de la main. Si on ignore un mot étranger, on peut se taire *(faire)* comprendre par des signes.

2. Monsieur Leclerc est fort, il a de la force il est robuste. André Caron fait des courses de dix ou quinze kilomètres, il est résistant. Madame Leclerc coud; elle est adroite; si elle était maladroite, le travail serait mal fait. La robe va bien; Madame Leclerc est habile; la fillette veut coudre aussi; elle a encore des gestes gauches.

3. Cette *(Cet)* âme *(homme)* à *(a)* une jambe plus courte que l'autre; il est boiteux; il boite de la jambe gauche; un accident l'as *(a)* rendu infirme. Les mutilés ont perdu un bas *(bras)*, une jambe ou un oeil dans un accident.

4. Le maître arrive. Les enfants se lèvent. Il *(ils)* reste *(restent)* debout. Le maître Guy: *(crie)* "assis". Dans le fonds *(fond)*, quelque *(quelques)* élève *(élèves)* n' ont pas entendu. Le maître répète: "asseyez vous" ... "acier *(assied)* toi, Daniel".

Figure 5: An example of a dictated text. The errors are underlined and followed by the correct wording inside parentheses.

Correct Sentence: se sont portés vers les problèmes relatifs
Recognized Sentence:

se	**sont**	**portés**	air	**les**	programme	**relatifs**
ce	son	porter	faire	lié	**problèmes**	relatif
ceux	sons	portées	heure	clé	problème	
CEE	soit		**vers**	clés		
seul	sans		...	mes		
...	ont			...		
	...					

(have been conducted on the problems related ...)

Figure 6: Sample output for the report dictation task. Although the top candidate string contains two errors, they are corrected by the language model.

Liaison:

Word string	*Phoneme string*	
des	/de/	(some)
amis	/ami/	(friends)
des amis	/dezami/	(some friends)
bon	/bO/	(good)
bon ami	/bcnami/	(good friend)
petit ami	/pxtitami/	(boy friend)
petits amis	/pxtizami/	(boy friends)

Apostrophe:

Word string	*Written form*	*Phoneme string*	
le ami	l'ami	/lami/	(the friend)
de ami	d'ami	/dami/	(from a friend)

Figure 7: Examples of liaison and apostrophe in French.

4.4 Discussion

Although isolated-word dictation helps to constrain the recognition task by removing the problem of finding the word boundaries, other problems are introduced. For example, it is not evident what to do about the liaison often made at word junctures. One possibility is to not pronounce the liaison at all, however, the resulting speech sounds very unnatural. Another option is to pronounce the liaison at the beginning of the following word, but this increases the size of the vocabulary, as all the possible liaisons at the beginning of the word must be allowed. A third possibility is to pronounce the liaison as a separate word, thus saying three words instead of two. This pronounciation of the liaison in isolation is very difficult since it is so unnatural. Another approach has been to dictate with isolated syllables instead of isolated words[38]. While this provides a more natural way to pronounce the liaison at the start of a syllable, the resulting task is still unnatural for the speaker.

A similar problem arises in that the vowel at the end of some words can be omitted when the next word begins with a vowel. In the resulting orthographic form, the vowel is replaced with an apostrophe, and the space separating the words is removed. The word sequence "le ami" thus becomes "l'ami". In pronouncing these words there are several options: The first one is to say the first word as if had not been modified, followed by the second word. The second option, which is to pronounce the words together as one word, has the unfortunate effect of greatly enlarging the size of the vocabulary. A third option is to say a sequence of three words, verbalizing the word "apostrophe" in the middle of the two other words. Some examples of these problems are given in Figure 7.

The problems associated with how to pronounce the liaisons and apostrophes in isolated word dictation emphasize the need for continuous dictation in French. While continuous dictation avoids these problems on the part of the speaker, they still remain for the recognizer, and increase its complexity.

5 Continuous-speech speech-to-text conversion

Our current efforts focus on speech-to-text conversion of continuously spoken sentences, from any speaker, for very large vocabularies (eventually, unlimited). This is a large departure from the approach taken in the previous section, where the task was speaker-dependent, isolated-word, and for smaller size vocabularies. Because of the ambitiousness of the task, the system should be both independent of the speaker and the vocabulary. To this extent, a phone-based approach is being used, where phone-like units are trained with data from a large number of speakers. In the next subsection some early work in phoneme-to-grapheme conversion from text is described. After a presentation of the corpus used for this work, the remainder of this section is devoted to current projects in phonetic and word recognition.

5.1 Phoneme-to-grapheme conversion from text

In light of the problems encountered in phoneme-to-grapheme conversion for continuous error-free phoneme strings, the use of a natural language syntactic parser[2] was explored. This work made use of a full-form dictionary containing almost 162,900 forms, derived from a 22,000 word base-form dictionary. Each graphemic word was converted into its phonemic form by using the automatic grapheme-to-phoneme conversion software designed at LIMSI[43]. The dictionary also includes other information such as the grammatical category of each word, its gender and number for the substantives and adjectives, the mode, time, person, group, transitivity, and root for the verb.

A positional syntax specified by a 3D frequential matrix giving the frequency of the succession of three grammatical categories was used[2]. (This kind of model is now commonly known as a trigram language model.) On the basis of linguistic knowledge and experimentation 150 grammatical categories were chosen.

The phoneme-to-grapheme conversion was tested on a 1800 word text, where the phonetic representation for each word was obtained using the same grapheme-to-phoneme conversion software as was used to represent the lexicon. Liaisons were not taken into account, though the punctuation markers were retained. All possible segmentations of the phonemic string were filtered by the trigram model. When several possibilities remained, the one with the smallest number of words was kept. The error rate was less than 5% on the 1,800 word test. The most common errors were:

- **singular/plural errors (36%)**, some of them being impossible to distinguish:
  plans/plan d'exécution (maps/map for execution)
  demande/demandes de permis (request/requests for permission)
- **homophones (17%)**:
  plan/plant (map/plant) heures/heurts (hours/collisions)
  ère/air/erre/hère/aire (era/air/wanders/wretch/area)
- **syntax parsing errors (17%)**:
  les baisses ont équipé / les baies sont équipées (the falls have equiped / the windows are equiped)
  et celles situées / et sels situés (and those situated / and salts situated)
- **number for posterior adjectives (13%)**:
  périmètre de protection des monuments historique/historiques (area of protection of historical/historical monuments)

Further work in phoneme-to-grapheme conversion was done as part of the ESPRIT project 291/860 on the Linguistic Analysis of the European Languages. An important part of the project was the building of a language model for 7 different languages (Italian, French, Dutch, Spanish, Greek, and German). A statistical approach was taken using bigram and trigram models on grammatical categories, similar to that developed at LIMSI.

The main results of this project were to provide statistics on phoneme clusters, grapheme-to-phoneme and phoneme-to-grapheme conversion software, language models and syntactic parsers. These elements were integrated using a blackboard structure and an attempt was made to assess the "quality" or "difficulty" of each language[6, 51].

On this last issue, some interesting results have been found in a study of phoneme-to-grapheme conversion for a lexicon of about 10,000 entries. One measure was the number of context-dependent rewrite rules necessary for phoneme-to-grapheme conversion. Table 7 shows that for Italian, a set of 67 rules is able to transcribe the phonemic form into the graphemic form with only 0.5% of the generated graphemic words not existing in the language, and 0.5% graphemic words unable to be transcribed. In contrast, for French, 98% of the words generated by a set of 586 rules do not exist in the vocabulary, and 30% of the vocabulary words are missing in the resulting graphemic cohorts. While this result is clearly highly dependent on the quality of the rules, it seems obvious that the Italian language will require less linguistic processing than the French language in order to translate a phonemic string.

Language	*# Rules*	*# Graphemic words/ phonemic words*	*% Over generation*	*% Under generation*
Dutch	289	6	90	20
English	530	10	90	6
French	586	250	98	30
German	551	400	99	10
Greek	394	100	100	2.5
Italian	67	1	0.5	0.5
Spanish	845	1	7	6

Table 7: Phoneme-to-grapheme translation for 7 European languages.

5.2 Database

For continuous speech recognition a portion of the BREF corpus is used. BREF is a large read-speech corpus, containing over 100 hours of speech material, from 120 speakers. The text materials were selected verbatim from the French newspaper *Le Monde*, so as to provide a large vocabulary (over 20,000 words) and a wide range of phonetic environments[18]. Containing 1115 distinct diphones and over 17,500 triphones, BREF can be used to train vocabulary-independent (VI) phonetic models. Hon and Lee[20] concluded that for VI recognition, the coverage of triphones is crucial. Separate text materials, with similar distributional properties were selected for training, development test, and evaluation purposes. The selected texts consist of 18 "all phoneme" sentences, and approximately 840 paragraphs, 3300 short sentences (average 12.4 words/sentence), and 3800 longer sentences (average 21 words/sentence). The "all phoneme" sentences contain all 35 phones given in Table 1. More

details of the distributional properties of the selected text subsets can be found in [18].

Each of 80 speakers read approximately 10,000 words (about 650 sentences) of text, and an additional 40 speakers each read about half that amount. The speakers, chosen from a subject pool of over 250 persons in the Paris area, were paid for their participation. Potential speakers were given a short reading test, containing selected sentences from *Le Monde* representative of the type of material to be recorded[30] and those judged to be incapable of the task were not used as subjects. The recordings were made in stereo in a sound-isolated room, and were monitored to assure the contents. Thus far, 80 training, 20 test, and 20 evaluation speakers have been recorded. There are 55 male and 65 female speakers. The speakers' ages range from 18 to 73 years, with 75% between the ages of 20 and 40 years. More details about the BREF corpus can be found in [30].

In these experiments approximately 4 hours and 20 minutes of speech material are used for training. This represents 2770 sentences from 57 speakers (28 male, 29 female). The test data consisted of 109 sentences from 19 speakers (10 male, 9 female). The test text material is distinct from the training texts, and the test speech data contain 7635 phone segments.

Phonemic transcriptions of these utterances were automatically generated and verified[15]. The procedure for providing a time-aligned broad phonetic transcription for an utterance has two steps. First, a text-to-phoneme module[43] generates the phone sequence from the text prompt. Since the automatic phone sequence generation can not always accurately predict what the speaker said, the transcriptions must be verified. The most common errors in translation occur with foreign words and names, and acronyms. Other mispredictions arise in the reading of dates: for example the year "1972" may be spoken as "mille neuf cent soixante-douze" or as "dix neuf cent soixante-douze." In the second step, the phone sequence is aligned with the speech signal using Viterbi segmentation.

5.3 Phone Recognition

In this section some experiments with phone recognition are described. Evaluating phonetic recognition is important for several reasons. Primarily, the demands of vocabulary-independent, speaker-independent continuous speech recognition require an approach based on subword, often, phone-like units. Clearly, the better these phone models (or acoustic models) are, the better the performance of the entire system will be. Only considering word recognition performance, particularly when word-based grammars are used, can mask problems that stem from the acoustic level. Phone recognition is also useful in determining pronunciation errors in the lexicon and alternate pronunciations that need to be included in the lexicon. Finally, phone recognition is shown to be effective for language identification and for speaker identification[25, 26, 27].

5.3.1 System Description

The baseline phone recognizer uses a set of 35 context-independent (CI) phone models. Each model is a 3-state left-to-right HMM with Gaussian mixture observation densities. The covariance matrices of all the Gaussians components are diagonal. The 16 kHz speech was downsampled by 2 and a 26-dimensional feature vector was computed every 10 ms. The feature vector is composed of 13 cepstrum coefficients and 13 differential cepstrum coefficients. Duration is modeled with a gamma distribution per phone model. As proposed by Rabiner et al.[46], the HMM and duration parameters are estimated separately and combined in the recognition process for the Viterbi search. Maximum likelihood estimators were used for the HMM parameters[21] and moment estimators for the gamma distributions.

Condition	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Acc.*
0-gram	64.0	23.7	12.3	3.0	61.0
1-gram	66.4	22.0	11.6	3.0	63.4
2-gram	70.2	20.3	9.4	3.1	67.1

Table 8: Phone recognition results for 35 CI models.

For CI models, the overall Markov chain is simply obtained by allowing all possible connections between the 35 phone HMMs (i.e. 1225 connections). For the transition probabilities either constant (1/35), 1-gram, or 2-gram probabilities were used. The resulting ergodic HMM has 103 states and about 170,000 parameters.

In the case of context-dependent (CD) models, the phone HMMs are connected through null states representing all the possible diphones. These null states, which do not emit any observation, are used to merge all the transitions corresponding to the same diphone, thus reducing the number of connections to a more manageable value. (The fourth order (n^4) becomes a cubic form). With 428 CD models, the resulting HMM includes 1294 non-null states and has about 1,070,00 parameters.

5.3.2 Evaluation

Table 8 gives recognition results using 35 CI phone models with 16 mixture components. Silence segments were not included in the computation of the phone accuracy. Results are given for different phone language models with a duration model. The improvement obtained by including the duration model is relatively small, on the order of 0.3% to 0.8%, probably in part due to the wide variation in phone durations across contexts and speakers. Each additional order in the language model adds about 3% to the phone accuracy. The best phone accuracy is 67.1% with the 2-gram language model.

Condition	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Acc.*
0-gram	75.3	17.8	6.9	4.1	71.2
1-gram	76.0	17.5	6.4	4.2	71.9
2-gram	77.8	16.5	5.7	3.6	74.2
8kHz, 32g, $\Delta\Delta$	81.7	13.7	4.6	3.0	78.7

Table 9: Phone recognition results for 428 CD models.

Table 9 gives recognition results using a set of 428 CD phone models[49] with 16 mixture components. The modeled contexts were automatically selected based on their frequencies in the training data. This model set is essentially composed of right-context phone models, with only one-fourth of the models being triphone models. Less than 2% of the triphones found in the training data can be modeled in full. In choosing to model right contexts over left contexts, a preference is given to modeling anticipatory coarticulation over perservatory coarticulation.

The phone accuracy with a phone bigram is 74.2%. The use of CD models reduces the errors by 22% (comparing the CI and CD models with the phone bigram), which is less than the 27% error reduction reported by Lee and Hon[32] for English. There are several factors that may account for this difference. Most importantly, Lee and Hon[32] compare 1450 right-CD models to 39 CI models, whereas in this study only 428 contexts were modeled. In addition, the baseline recognition accuracy reported by Lee and Hon is 53.3% with a bigram

language model, compared to our baseline phone accuracy of 67.1%. In these experiments using as many as 2100 CD models did not significantly reduce the error rate.

The use of a different signal analysis (8kHz MFCC), and additional parameters (32 Gaussians per mixture, the second derivative of the cepstrum ($\Delta\Delta$)) and sex-dependent models improves the phone accuracy to 78.7% as shown in the last entry in the table[28].

5.3.3 Error Analysis

Confusion pair	*% Subs.*
e $\rightarrow$ E	4.2
E $\rightarrow$ e	3.8
a $\rightarrow$ E	4.2
E $\rightarrow$ a	1.8
n $\rightarrow$ m	1.8
y $\rightarrow$ i	1.8

Table 10: The most common substitutions with 428 models.

The most recognition errors occurred for the phones: /E/ 8.1%, /a/ 7.6%, /e/ 7.2%, /c/ 4.9%, /t/ 4.3%,and /x/ 4.2%, accounting for almost 40% of the substitution errors. Of these phones only /c/ and /E/ have high phone error rates of about 40%. Table 10 shows the most frequent substitutions made by the recognizer. The two most common confusions are reciprocal confusions between /e/ and /E/ and between /E/ and /a/. Together these account for 13% of the confusions. Many speakers do not make a clear distinction between the phones /E/ and /e/ when they occur word-internally, which may account for their high confusability. The high number of errors for /a/ are probably due to the large amount of variability of /a/ observed in different contexts.

14% of the insertions are /r/, followed by 11% for /l/. These two phones also are deleted the most: 13% of the deletions are /l/ and 11% /r/. Although /l/ and /r/ account for many of the insertion and deletion errors, the overall error rate for these phones are relatively low, 11% and 7%, respectively. Improved performance on these phones may be achieved by modeling more contexts and by improving their duration models.

5.3.4 Language Identification

Another application for which phonetic recognition is used is language identification, which could be a component of a multilingual speech-to-text system. The basic idea is to process in parallel the unknown incoming speech by different sets of phone models for each of the languages under consideration, and to choose the language associated with the model set providing the highest likelihood. The language-dependent models are trained from similar-style corpora, BREF for French and WSJ0 for English, both containing read newspaper texts and similar size vocabularies[18, 30, 41]. A set of SI CI phone models were built for each language, with 35 models for French and 46 models for English.[16, 29] Each phone model has 32 gaussians per mixture, and no duration model. In order to minimize influences due to the use of different microphones and recording conditions a 4 kHz bandwidth is used. The training data for French were 2770 BREF sentences from 57 speakers and for English the WSJ0 SI-84 data containing 7240 sentences from 84 speakers.

Language identification accuracies are given in Table 11 with phonotactic constraints provided by a phone bigram. Language identification error rates are given for the 4 test

corpora, WSJ and TIMIT for English[41, 11], and BREF and BDSONS for French[18, 30, 7], as a function of the duration of the speech signal. Approximately 100ms of silence are included at the beginning and end of each utterance (the initial and final silences were automatically removed based on HMM segmentation), so as to be able to compare language identification as a function of duration without biases due to long initial silences. The test data for WSJ0 consist of 100 sentences, the first 10 sentences for each of the 10 speakers (5m/5f) in the Feb92-si5Knvp (speaker-independent, 5K, non-verbalized punctuation) test data. For TIMIT, the 192 sentences in the "coretest" set containing 8 sentences from each of 24 speakers (16m/8f) were used. The BREF test data consist of 130 sentences from 20 speakers (10m/10f) and for BDSONS the data are comprised of 121 sentences from 11 speakers (5m/6f).

Test Corpus	*# of sents*	*Error rate vs. Duration*					
		0.4s	*0.8s*	*1.2s*	*1.6s*	*2.0s*	*2.4s*
WSJ	100	5.0	3.0	1.0	2.0	1.0	1.0
TIMIT	192	9.4	5.7	2.6	2.1	0.5	0
BREF	130	8.5	1.5	0.8	0	0.8	0.8
BDSONS	121	7.4	2.5	2.5	1.7	0.8	0
Overall	543	7.9	3.5	1.8	1.5	0.7	0.4

Table 11: Language identification error rates as a function of duration and language with phonotactic constraints provided by a phone bigram. (The duration includes 100ms of silence.)

The overall French/English language identification error is less than 1% with 2s of speech. It can be seen in Table 11 that while WSJ sentences are more easily identified as English for short durations, errors persist longer in these sentences than for TIMIT. In contrast for French, BDSONS data are better identified than BREF with 400ms of signal, perhaps because the sentences are phonetically balanced. For longer durations, BREF is slightly better identified than BDSONS. Bearing in mind that the corpora were recorded under similar conditions, the performance demonstrated here shows that accurate task-independent, cross-corpus language identification can be achieved. Extensions of this work will include identification of other European languages.

5.3.5 Speaker Identification

The same approach has also been used for text-independent speaker identification[16, 29]. In this case a set of phone models were built for each speaker, and the unknown speech was recognized by all of the speakers models in parallel. The base acoustic models were the 35 CI BREF models, built using the training data from the 57 training speakers. These models were adapted to each of 65 speakers (including 8 new speakers) using only 8 of the training sentences, and 2 sentences were used for identification test. Using only one sentence per speaker for identification, there is one error, giving an identification accuracy of 99.2%. When 2 sentences are used all speakers are correctly identified.

Experiments for English used a set of 40 SI CI models trained on the 462 training speakers in the TIMIT corpus[11] as seed models to estimate 31-phone model sets for each of the 168 test speakers in TIMIT. Using 8 sentences (2 SA, 3 SX, and 3 SI) for adaptation resulted in 98.5% correct speaker identification using one sentence for identification and 100% identification if the likelihood over two sentences was used. Recently, high speaker identification rates using subsets of 100 to all 462 speakers from TIMIT have been reported[5, 40, 48].

A simple reduction in computation is gained by first determining the sex of the speaker by running in parallel SI male and female models. In experiments with this approach no cross-sex errors have ever occurred with the SI male/female models or with any of the SD models. Further computational reductions during recognition can be obtained by speaker clustering.

5.4 Word Recognition

Two types of implementation are usually considered to recognize words based on phone models. In the first solution, which can be called *integrated approach*, an HMM is generated for each word by concatenating the phone models according to the phone transcriptions. The word models are put together to represent the entire lexicon with one large HMM. The recognition process is then performed for example by using the Viterbi decoding algorithm. The second solution uses the output of the phone recognizer as an intermediary level of coding such that the lexical decoding is derived only from this output. Phonological rules may be included in the lexical decoding, or alternatively may be represented directly in the lexical entries. The phone recognizer output is usually a phone trellis including phone hypotheses for each of the associated speech segments and their corresponding likelihoods. If the first approach appears to offer a more optimal solution to the decoding problem by avoiding an intermediary coding, the second approach greatly reduces the computional requirements of the acoustic level which is independent of the lexicon size and allows lexical and language models to be developed and evaluated without interaction with the acoustic level.

The recognizer used in the experiments described in this section uses the integrated approach, a time-synchronous graph-search strategy. This one level implementation includes intra- and inter-word context-dependent (CD) phone models, intra- and inter-word phonological rules, phone duration models, gender-dependent models, and can be used with different types of language models including word-pair and bigram grammars[25, 17]. For this evaluation, liaison is represented in the lexicon as alternate pronunciations. Since this simplistic approach is not practical for large lexicons, other approaches, such as the use of phonological rules are being investigated to allow optional liaison. Phonological rules were shown to be effective in reducing the error rate for the DARPA Resource Management Task[25]. This latter solution has the advantage that it is not necessary to expand the lexical pronunciations which can have the undesired side-effect of overgeneralization. The environments in which liaison, and other such events, are allowed are specified by the phonological rules, which are used both for training and recognition.

5.4.1 System Description

For all the experiments the same set of 428 CD phone models already used for the phone recognition experiments are used. For the no-grammar case a phone tree is built from the lexicon in order to reduce the graph size. For the 10K lexicon the average number of phone nodes per word goes from 6.4 to 2.0 by using such a tree instead of a linear representation of each word, i.e. a 69% graph size reduction. For the word-pair and bigram grammars, a phone graph is first built by linking the word phone transcriptions according to the grammar, then, as for the no-grammar case, the phone graph is converted to a large HMM by replacing each phone node by the appropriate set of phone models and establishing the proper connections with the neighboring phones. A bigram-backoff[22] language model estimated on the text material from *Le Monde* is used for lexicons containing 5K and 20K words. In all cases, CD phone models are used for word juncture phones as well as for intra-word phones.

As an example, for the 3K lexicon, the average number of instanciations of each phone model is 98 for the no-grammar case and 96 for the word-pair grammar. Considering the differences in the network representation, these numbers are surprisingly similar. The memory used with the word-pair grammar is 24 Mb, compared to 17 Mb for the no-grammar case. Most of the difference is due to the interword connections for the word-pair grammar.

5.5 Evaluation

Lexicon	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Acc.*
1K	73.4	20.9	5.8	4.2	69.2
3K	66.5	27.5	6.0	5.3	61.2
5K	61.4	32.0	6.6	5.9	55.6
10K	55.4	36.9	7.7	6.4	49.0

Table 12: VI word recognition results (no grammar).

Vocabulary-independent word recognition experiments were run using four different lexicons. The smaller lexicon (1K lexicon) contains 1139 orthographic words, only those words found in the test sentences. The 3K lexicon contains all the words found in the training and test sentences, a total of 2716 words. The 5K and 10K lexicons include all the words in the test data complemented with the most common words in the original text. These two lexicons contain respectively 4863 and 10511 words. Alternate pronunciations increase the number of phonemic forms in the lexicon by about 10%. The word recognition results with no grammar are given in Table 12. Since no grammar is used, single word homophone confusions are not counted as errors.

As discussed in Section 3, homophones present a large problem for French. If the homophone errors are included the phone accuracies drop by about 10%. A lexical study with 300,000 words found that there can be over 30 words with the same pronunciation. In the *Le Monde* text corpus of 4.2 million words, there were 92,185 orthographically distinct words, but only 63,981 phonemically distinct words, giving a homophone rate of about 30%. In the 1K and 3K lexicons the homophone rate is lower, on the order of 15%. The "worst-case" homophone in the 3K lexicon is for the phonemic word /sA/, which may correspond to any of the 7 following orthographic words: *100, cent, cents, s'en, sang, sans, sent.*

While the large number of word homophones in French presents its problems, more complicated homophone problems exist, where sequences of words form homophones. The example in Figure 8 shows some of the homophones for the phonetic sequence /parle/ for the words in the 3K lexicon. These multiple word homophones account for a few percent of the errors in Table 12. In fluent speech, the problems are more complicated as illustrated by Figure 9. While nominally the phonetic transcription of the word "adolescence" is /adclEsAs/, the realized pronunciation is /adxlEsAs/, having the given homophones.

Lexicon	*Perp.*	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Acc.*
1K	100	90.1	8.5	1.4	2.2	87.9
3K	160	88.2	10.2	1.5	2.2	86.1

Table 13: Word recognition results with a word-pair grammar.

Phonetic transcription:	/p a r l e/
Word candidates:	parler
	parlé
	parlée
	par les
	part les
	parle es
	parlent es
	parle et
	parlent et

Figure 8: An example of a multiple word homophone.

Phonetic transcription:	/a d x l E s A s/
Word candidates:	adolescence
	a de les sans
	a de l'essence

Figure 9: An example of a homophone caused by vowel reduction.

Word recognition results using a word-pair grammar derived on the entire 4.2 million word text of *Le Monde* are given in Table 13 for the 1K and 3K lexicons. Since a grammar is used homophones are counted as errors. The use of the word-pair grammar reduces the perplexities to 100 for the 1K lexicon and 160 for the 3K lexicon, and reduces the error rate by almost 60%. In addition, the drop in performance observed by increasing the lexicon size is smaller than for the no grammar case, as is expected given that the perplexity is not proportional to the size of the lexicon.

Two vocabularies have been used for recognition experiments with bigram-backoff language model, containing only the 5,000 and 20,000 most common words in the *Le Monde* texts. The test data consist of 100 sentences for each vocabulary size, with perplexities of 122 for the 5K sentences and 205 for the 20K sentences. These are not the same test data as used in the no-grammar and word-pair grammar conditions since the test texts were selected so that all the words were found in the respective lexicons. Word recognition results using the same 428 CD models and the bigram-backoff language model are shown in Table 14. The acoustic processing and feature vector are the same as used for the improved performance phone recognizer. The word error is 14.5% for the 5K lexicon and 18.3% for the 20K lexicon. More details of the experimental conditions and results can be found in [17].

6 Summary

In this paper an overview of the research at LIMSI in the area of speech-to-text conversion has been given. Research projects in this domain have been pursued since the 1970's. The projects include phoneme-to-grapheme conversion of ideal and errorful strings, isolated-word speech recognition, and continuous speech recognition. Throughout problems that are specific

Lexicon	*Perp.*	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Acc.*
5K	122	87.1	10.3	2.6	1.7	85.5
20K	205	84.6	12.8	2.6	2.9	81.7

Table 14: Word recognition results on the BREF80 corpus with a probabilistic grammar (2-grams) estimated on *Le Monde* text data. (5K: 5000 word lexicon, 20K: 20,000 word lexicon).

to French have been highlighted.

Speech-to-text conversion of French presents difficulties different from those found in English. One major problem is dealing with the large number of homophones. Lexical studies indicate a single-word homophone rate of about 30%[18, 15]. Comparative homophone rates for English are roughly 3% for DARPA Resource Managment lexicon[42] and less than 2% for the DARPA TIMIT lexicon[31, 11]. The main problem comes from the conjugation of verbs, and markers for plural (s) and feminine (e) at the end of some classes of words (past participles, some adjectives, etc.) which are not pronounced.

In part due to the high homophone rate, the segmentation of even error-free continuous phoneme strings into words seems to be especially difficult in French. For example, the simple sentence containing 9 phonemes, "J'ai mal au pied." (My foot hurts.), has more than 32,000 possible transcriptions at the lexical level with a 162,900 word full-form lexicon. Even using phonological rules, syntax, and semantics two sentences remained which require a pragmatic analysis to determine the correct graphemic transcription.

Liaison is another problem that must be dealt with. This word-juncture event is never pronounced in isolation and is optionally pronounced in continuous speech. How to pronounce the liaison is a problem particular to isolated word dictation, that is solved in continuous speech. The problem is even more complicated in that sometimes this optional liaison is the only indication to distinguish between the singular and plural forms of a word or phrase. Being optional, liaison increases the number of inter-word connections. The formalism demonstrated in the framework of the RM task[25] is being used to handle this problem. This formalism uses phonological rules to account for alternate pronunciations and to handle cross-word coarticulation.

Another problem which can be handled similarly with the use of phonological rules is the optional pronunciation of mute-e: Certain words may be pronounced with either 2 or 3 syllables; the schwa in short function words may be completely deleted; and the final usually silent mute-e at ends of words may be pronounced. Another problem concerns the apostrophe, where the final vowel of certain words can be deleted when the next word begins with a vowel.

Our most recent work focuses on developing phone-based speech recognizers that are task, speaker and vocabulary independent so as to be easily adapted to various applications. The recognizer described here was evaluated at both the phone and word levels. A set of 428 context-dependent models were trained on speech taken from 57 speakers in the BREF corpus. These were tested on 109 sentences taken from a new 19 speakers. The resulting phone accuracy was 78.7%, with phonotactic constraints given by a phone bigram. The phone recognition results are encouraging and are somewhat superior to those reported for English[32, 47, 28]. This may be simply because French has a smaller number of phonemes, or that the phonemes are less variable due to context.

It is our opinion that it is important to evaluate the quality of the acoustic models, and

that phone recognition provides a relevant means for doing so. This has the added benefit that the recognized phone string can be used to understand errors in word recognition, and problems with the lexical representation. Phone recognition has also been found to be powerful for language identification and speaker identification. The approach is straight forward. Multiple model sets are run in parallel, and the language (or speaker) is identified as that language (or speaker) associated with the model having the highest likelihood. Experiments in language identification show that with 2s of speech the language is correctly identified as English or French with 99% accuracy. Speaker identification experiments with TIMIT have a speaker-identification rate of 98.5%, comparing each speaker to models from all 168 test speakers using 1 utterance per speaker, and 100% correct if two utterances are used.

Word recognition for BREF was evaluated on lexicons ranging from 1000 to 20,000 words, for the no-grammar case, with a word-pair grammar, and with a bigram-backoff grammar. For the no-grammar case the word accuracy was 69.2% with the 1K lexicon and dropped to 49% with the 10K lexicon. With a word-pair grammar the word accuracy was 87.9% and 86.1% respectively for the 1K and 3K grammars. The word accuracies on a different set of test sentences was 85.5% for the 5K vocabulary and 81.7% for the 20K vocabulary.

References

[1] G. Adda (1987), *Reconnaissance de Grands Vocabulaires: Une étude Syntaxique et Lexicale*, Thèse de Docteur-Ingénieur, Université Paris XI, December, 1987.

[2] A. Andreewski, J.P. Binquet, F. Debili, C. Fluhr, Y. Hlal, J.S. Liénard, J. Mariani, B. Pouderoux (1979), "Les dictionnaires en forme complète, et leur utilisation dans la transformation lexicale et syntaxique de chaînes phonétiques correctes," 10èmes "Journées d'Etudes sur la Parole" du "Groupement des Acousticiens de Langue Française," Grenoble, May 1979, pp. 285-294 .

[3] L.R. Bahl, R. Bakis, P.S. Cohen, F. Jelinek, B.L. Lewis, R.L. Mercer (1978), "Recognition of a Continuously Read Natural Corpus," *Proc. IEEE ICASSP-78*, Tulsa, AZ, April 1978, pp. 422-425.

[4] D. Bellilty and A. Lund (1984), "Conversion phonèmes-graphèmes de suites phonétiques entachées d'erreurs," LIMSI internal report, July, 1984.

[5] Y. Bennani (1992), "Speaker Identification through a Modular Connectionist Architecture: Evaluation on the TIMIT Database," *Proc. ICSLP-92*, Banff, Canada, Vol. 1, pp. 607-610.

[6] L. Boves, M. Refice *et al.* (1987), "The Linguistic Processor in a Multi-Lingual Text-to-Speech and Speech -to -Text System," *European Conference on Speech Technology*, Edinburgh, September 1987, pp. 385-388.

[7] R. Carré, R. Descout, M. Eskénazi, J. Mariani, M. Rossi, "The French language database: defining, planning, and recording a large database," *ICASSP-84*.

[8] A.M. Derouault (1985), *Modélisation d'une langue naturelle pour la désambiguation des chaînes phonétiques*, Thèse de Doctorat d'Etat, Univ. Paris VII, April 1985.

[9] V. Digalakis, M. Ostendorf, J.R. Rohkicek (1990), "Fast Search Algorithms for Connected Phone Recognition Using the Stochastic Segment Model," *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, June 1990, pp. 173-178.

[10] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," Proc. DARPA Speech Recog. Workshop, 1986.

[11] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren (1993), "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM" (printed documentation for NIST Speech Disc 1-1.1), NTIS order number PB91-100354.

[12] J.L. Gauvain (1986), "A Syllable-Based Isolated Word Recognition Experiment," *Proc. IEEE ICASSP-86*, Tokyo, Japan, April 1986, pp. 57-60.

[13] J.L. Gauvain (1990), "Le système de reconnaisance *AMADEUS*: Principe et algorithmes," LIMSI internal report, June 1990.

[14] J.L. Gauvain and J.J. Gangolf (1983), "Terminal integrates speech recognition and text-to-speech synthesis", *Speech Technology*, Sept-Oct 1983.

[15] J.L. Gauvain and L.F. Lamel (1992), "Speaker-Independent Phone Recognition Using BREF," *Proc. DARPA Speech and Natural Language Workshop*, Arden House, NY, Feb. 1992.

[16] J.L. Gauvain, L. Lamel, "Identification of Non-Linguistic Speech Features ," *ARPA Workshop on Human Language Technology*, Plainsboro, NJ, March, 1993.

[17] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, " Speaker-Independent Continuous Speech Dictation," *Proc. EUROSPEECH-93*, Berlin, Germany, Sept. 1993.

[18] J.-L. Gauvain, L.F. Lamel, M. Eskénazi (1990), "Design Considerations and Text Selection for BREF, a large French read-speech corpus," *Proc. ICSLP-90*, Kobe, Japan, Nov. 1990, pp. 1097-2000.

[19] J.L. Gauvain and J. Mariani (1982), "A Method for Connected Word Recognition and Word Spotting on a Microprocessor", *Proc. IEEE ICASSP-82*, Paris, France, May 1982, pp. 891-894.

[20] H.-W. Hon and K.-F. Lee (1990),"On Vocabulary-Independent Speech Modeling," *Proc. IEEE ICASSP-90*, pp. 725-728.

[21] B. H. Juang (1985), "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains", *AT&T Technical Journal*, Vol. 64, No. 6, July-August 1985.

[22] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. ASSP*, **35**(3), 1987.

[23] M.H. Kuhn and H.H. Tomaschewski (1983), "Improvements in Isolated Word Recognition," *IEEE Trans. on ASSP*, Vol. 31, N. 1, February 1983, pp. 157-167.

[24] L.F. Lamel and J.-L. Gauvain (1992), "Experiments on Speaker-Independent Phone Recognition Using BREF," *Proc. IEEE ICASSP-92*, San Francisco, CA, Vol. S1, pp. 557-560.

[25] L.F. Lamel and J.-L. Gauvain (1992), "Large-Vocabulary Speech Recognition at LIMSI," presented at the final review of the *DARPA Artificial Neural Network Technology (ANNT) Speech Program*, Stanford, CA, Sept. 21-22.

[26] L.F. Lamel and J.-L. Gauvain (1992), "Multi-lingual Speech Recognition at LIMSI," Presented at the 1st International Workshop of Speech Translation, Warden, Germany, Oct. 18-20.

[27] L.F. Lamel and J.-L. Gauvain (1993), "Cross-Lingual Experiments with Phone Recognition," *Proc. IEEE ICASSP-93*, Minneapolis, MN.

[28] L.F. Lamel, J.L. Gauvain, "High Performance Speaker-Independent Phone Recognition Using CDHMM," *Proc. EUROSPEECH-93*, Berlin, Germany, Sept. 1993.

[29] L. Lamel, J.L. Gauvain, "Identifying Non-Linguistic Speech Features," *Proc. EUROSPEECH-93*, Berlin, Germany, Sept. 1993.

[30] L.F. Lamel, J.-L. Gauvain, M. Eskénazi (1991), "BREF, a Large Vocabulary Spoken Corpus for French," *Proc. EUROSPEECH-91*, Genoa, Italy, pp. 505-508.

[31] L.F. Lamel, R.H. Kassel, and S. Seneff, "Speech Database Development: Design and analysis of the acoustic-phonetic corpus," *Proc. DARPA Speech Recognition Workshop*, 1986.

[32] K.-F. Lee, H.-W. Hon (1989), "Speaker-Independent Phone Recognition Using Hidden Markov Models," *Proc. IEEE Trans. ASSP*, Vol. 37, No. 11, pp. 1641-1989.

[33] S.E. Levinson, M.Y. Liberman,A. Ljolje, L.G. Miller (1989), "Speaker Independent Phonetic Transcription of Fluent Speech for Large Vocabulary Speech Recognition," *Proc. IEEE ICASSP-89*, Glascow, Scotland, May 1989, pp. 441-444.

[34] J. Mariani (1977), *Contribution à la Reconnaissance de la Parole Continue utilisant la notion de Spectre Différentiel*, Thèse de Docteur-Ingénieur, Université Paris VI.

[35] J. Mariani (1981), "Reconnaissance de parole continue par diphonèmes," Séminaire du "Groupement des Acousticiens de Langue Française:" "Processus d'encodage et de décodage phonétique," Toulouse, September, 1981.

[36] J. Mariani (1987), "HAMLET: A Prototype of a Voice-Activated Typewriter," *Proc. European Conference on Speech Technology*, Edinburgh, September, 1987.

[37] B. Merialdo, A.-M. Derouault, S. Soudoplatoff (1986), "Phoneme Classification using Markov Models," *Proc. IEEE ICASSP-86*, Tokyo, Japan, April 1986, pp. 2759-2762.

[38] B. Merialdo (1987), "Speech Recognition with Very Large Vocabulary," *Proc. IEEE ICASSP-87*, Dallas, TX, pp. 364-367.

[39] B. Merialdo (1988), "Phonetic Recognition Using Hidden Markov Models and Maximum Mutual Information Training," *Proc. IEEE ICASSP-88*, New York, NY, pp. 111-114.

[40] C. Montacié and J.L. Le Floch (1992), "AR-Vector Models for Free-Text Speaker Recognition," *Proc. ICSLP-92*, Banff, Canada, Vol. 1, pp. 611-614

[41] D. Paul and J. Baker (1992), "The Design for the Wall Street Journal-based CSR Corpus," *Proc. DARPA Speech and Natural Language Workshop*, Arden House, Feb. 1992, pp. 357-362.

[42] P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett (1988), "The DARPA 1000-word Resource Management Database for Continuous Speech Recognition," *Proc. IEEE ICASSP-88*, New York, NY, pp. 651-654.

[43] B. Prouts (1980),*Contribution à la synthèse de la parole à partir du texte: Transcription graphème-phonème en temps réel sur microprocesseur*, Thèse de docteur-ingénieur, Université Paris XI, November, 1980.

[44] G. Quénot, J.L. Gauvain, J.J. Gangolf, J. Mariani (1986), "A dynamic time warp VLSI processor for continuous speech recognition", *Proc. IEEE ICASSP-86*, Tokyo, Japan, April 1986, pp. 1549-1552.

[45] G.M. Quénot, J.L. Gauvain, J.J. Gangolf, and J. Mariani (1989), "A Dynamic Programming Processor for Speech Recognition", *IEEE J. of Solid-State Circuits*, Vol. 24, No. 2,i April 1989, pp. 349-357.

[46] L.R. Rabiner, B.H. Juang, S.E. Levinson, M.M. Sondhi (1985), "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities," *AT&T Technical Journal*, 64(6), pp. 1211-1233, July-Aug. 1985.

[47] T. Robinson and F. Fallside (1991), "A recurrent error propogation network speech recognition system," *Computer Speech and Language*, Vol. 5, pp. 259-274.

[48] M. Savic, J. Sorenson (1992), "Phoneme Based Speaker Verification," *Proc. IEEE ICASSP-92*, San Francisco, CA, Vol. II, pp. 165-168.

[49] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, J. Makhoul (1985), "Context-dependent modeling for acoustic-phonetic recognition of continuous speech," *Proc. IEEE ICASSP-85*, Tampa, FL ,pp. 1205-1208.

[50] H. Singer and J.L. Gauvain, (1988) "Connected speech recognition using dissyllable segmentation,"*Fall meeting of the Acoust. Soc. of Japan.*

[51] V. Vittorelli (1987), "Linguistic Analysis of the European Languages," ESPRIT'87 Achievements and Impact, North-Holland, 1987, pp. 1358-1366.

[52] M. Webster (1964), *Pocket Dictionary*, computer readable form.